Semi-supervised Kernel Canonical Correlation Analysis with Application to Human fMRI

Matthew B. Blaschko[*,a], Jacquelyn A. Shelton[b], Andreas Bartels[c,d], Christoph H. Lampert[e], Arthur Gretton[f,g]

[a] *Department of Engineering Science, University of Oxford, United Kingdom*
[b] *Frankfurt Institute for Advanced Studies, Goethe Universität Frankfurt, Germany*
[c] *Centre for Integrative Neuroscience, Universität Tübingen, Germany*
[d] *Department of Neurophysiology, Max Planck Institute for Biological Cybernetics, Germany*
[e] *Institute of Science and Technology (IST), Austria*
[f] *Gatsby Unit, University College London, United Kingdom*
[g] *Department of Empirical Inference, Max Planck Institute for Biological Cybernetics, Germany*

Abstract

Kernel Canonical Correlation Analysis (KCCA) is a general technique for subspace learning that incorporates principal components analysis (PCA) and Fisher linear discriminant analysis (LDA) as special cases. By finding directions that maximize correlation, KCCA learns representations that are more closely tied to the underlying process that generates the data and can ignore high-variance noise directions. However, for data where acquisition in one or more modalities is expensive or otherwise limited, KCCA may suffer from small sample effects. We propose to use semi-supervised Laplacian regularization to utilize data that are present in only one modality. This approach is able to find highly correlated directions that also lie along the data manifold, resulting in a more robust estimate of correlated subspaces.

Functional magnetic resonance imaging (fMRI) acquired data are naturally amenable to subspace techniques as data are well aligned. fMRI data of the human brain are a particularly interesting candidate. In this study we implemented various supervised and semi-supervised versions of KCCA on human fMRI data, with regression to single and multivariate labels (corresponding to video content subjects viewed during the image acquisition). In each variate condition, the semi-supervised variants of KCCA performed better than the supervised variants, including a supervised variant with Laplacian regularization. We additionally analyze the weights learned by the regression in order to infer brain regions that are important to different types of visual processing.

Key words: Canonical Correlation Analysis, Semi-supervised Learning, fMRI

1. Introduction

Canonical correlation analysis (CCA) is a fundamental technique in statistics and dimensionality reduction that relies on paired data to learn directions that maximize correlation between the projected representations in each space [1]. It is readily kernelized (KCCA), enabling a straightforward non-linear generalization [2, 3, 4, 5]. Dimensionality reduction techniques that rely on only one modality are incapable of distinguishing semantically meaningless noise directions, and are not discriminative in nature. In contrast, KCCA is able to learn relevant directions by requiring that embedded data be correlated with embeddings of data in other modalities, and has been shown to increase class separability when compared to single modality dimensionality reduction [6].

While KCCA often gives superior results to dimensionality reduction techniques that work on a single modality, it does not directly estimate the data manifold in any given modality. Additionally, it is only able to utilize data for which correspondence is known to the other modalities. In order to more robustly learn the relevant directions in the feature space, we can modify our objective to favor directions that lie along the data

*Corresponding author

Email addresses: blaschko@robots.ox.ac.uk (Matthew B. Blaschko), shelton@fias.uni-frankfurt.de (Jacquelyn A. Shelton), andreas.bartels@tuebingen.mpg.de (Andreas Bartels), chl@ist.ac.at (Christoph H. Lampert), arthur.gretton@gmail.com (Arthur Gretton)

manifold. In this work, we describe a method that incorporates these two goals by employing semi-supervised Laplacian regularization [7]. This method gives an embedding of the data that makes use of the information between modalities, as well as the information within each single modality. By using Laplacian regularization, we are able to learn directions that tend to lie along the data manifold estimated from a much larger set of data [7]. This gives us greater confidence that the learned directions represent the underlying statistical structure of the data and that we have not been misled by small sample effects. We show experimentally that learning along the manifold results in increased performance, even in the fully supervised setting, in that the learned embeddings give better hold out correlations for a human fMRI task. Additionally, we show that the learned projection vectors are interpretable as representing brain regions that are implicated in the corresponding visual processing task.

Subspace methods have been applied to aligned image data in several classic computer vision applications. Perhaps most famously, Turk and Pentland applied principal components analysis (PCA) to aligned human faces, resulting in a basis that indicates the major variation in the training data [8, 9]. Belhumeur et al. extended this setting to apply Fisher linear discriminant analysis (LDA) in place of PCA, resulting in a subspace that discriminates between classes while ignoring high variance, indiscriminative directions [10]. When modelling visual data in this fashion, the data must meet the criterion of being reasonably well-aligned, otherwise translational biases are introduced. For this reason, images obtained via functional magnetic resonance imaging (fMRI), such as of human brains, are naturally suited to these techniques. Such fMRI data is obtained via an MRI scanner, which holds the subjects' heads immobile, thereby acquiring reasonably well-aligned data from the start. The data can be then further aligned using a suite of techniques, developed for neuroscience image analysis, e.g. Friston et al. [11]. This data processing pipeline is automatic and can be assumed to be available without significant additional human cost, making fMRI data a perfect candidate for the application of KCCA and related techniques. Several pattern recognition techniques have previously been applied to brain imaging data, including support vector machines, random forests, and Fisher LDA [12, 13, 14, 15, 16]. In Hardoon et al. [17], KCCA was applied to fMRI data from human subjects.

Although KCCA has been applied in many situations, including cross media information retrieval [5, 18], multi-modal data clustering [6], analysis of fMRI data [17], extraction of gene clusters [19], testing for independence [20, 21], and ICA [4], to our knowledge the only semi-supervised extension of the algorithm is due to the authors of this article [22]. Laplacian regularization is a common technique for semi-supervised learning [7, 23]. Cai et al. [24] have proposed a semi-supervised Fisher linear discriminant analysis algorithm based on Laplacian regularization, which we show in Section 3.4 to be a special case of the algorithm proposed here. We have recently proposed the use of Laplacian regularized ridge regression for semi-supervised fMRI analysis with resting state data [25]. de Sa et al. [26] have developed an algorithm for spectral clustering that is closely related to KCCA. Finally, temporal correlations were modeled in KCCA in Bießmann et al. [27]

In Section 2 we provide a review of kernel canonical correlation analysis, and in Section 3 we present in greater detail the semi-supervised Laplacian regularization introduced in Blaschko et al. [22]. We discuss its application to functional magnetic resonance imaging in Section 4, and we present in Section 5 empirical results showing improved performance compared to KCCA without Laplacian regularization both quantitatively and from a qualitative neuroscience perspective.

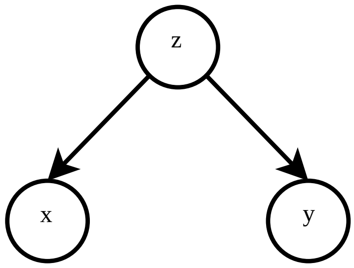
2. A Review of Kernel Canonical Correlation Analysis

2.1. Canonical Correlation Analysis

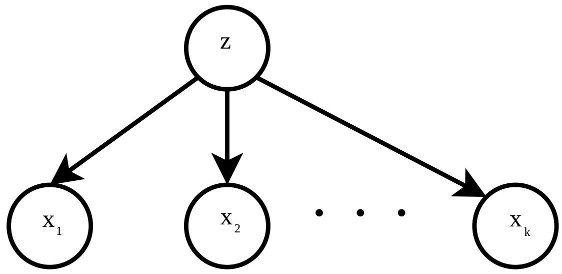
Canonical correlation analysis (CCA) utilizes datasets where samples are available in more than one modality. In the most simple case, this consists of paired data, i.e. data are present in only two modalities, but we also consider the more general case for which three or more modalities are present (Figure 1). CCA projects the data samples from each modality into a subspace such that the empirical correlation of the projected data is maximized [1]. Given a sample from a paired dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, CCA simultaneously finds directions w_x and w_y that maximize the correlation of the projections of x onto w_x with the projections of y onto w_y. This is expressed as

$$\max_{w_x, w_y} \frac{\hat{E}[\langle x - \mu_x, w_x\rangle\langle y - \mu_y, w_y\rangle]}{\sqrt{\hat{E}[\langle x - \mu_x, w_x\rangle^2]\hat{E}[\langle y - \mu_y, w_y\rangle^2]}} \tag{1}$$

where $\hat{E}$ denotes the empirical expectation, and μ_x and μ_y the empirical means in each of the modalities. We may view the general assumptions of CCA as being that



(a) Paired data.



(b) Multimodal data.

Figure 1: In a paired data set, there are two observed output variables, x and y that are generated by some underlying process z. This underlying process induces a dependence between x and y (a). When data are present in multiple modalities, an underlying process may induce dependence between all observed variables (b).

samples from $\mathcal{X}$ and $\mathcal{Y}$ are generated from some underlying process which induces a dependence between our paired samples (Figure 1).

We introduce the notation C to represent the covariance matrix of samples in $\mathcal{X} \times \mathcal{Y}$, and note that C decomposes into auto-covariance matrices, and cross-covariance matrices

$$C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \tag{2}$$

where C_{xx} and C_{yy} are auto-covariance matrices, and $C_{xy} = C_{yx}^T$ are cross covariance matrices. Using this notation, we may rewrite Equation (1) to obtain

$$\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x \; w_y^T C_{yy} w_y}}. \tag{3}$$

This Rayleigh quotient can be optimized as a generalized eigenvalue problem, or by decomposing the problem as described in Hardoon et al. [5].

In general, samples may be available in many modalities, $(\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k)$, yielding

$$C = \begin{pmatrix} C_{11} & C_{12} & \dots & C_{1k} \\ C_{21} & C_{22} & \dots & C_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ C_{k1} & C_{k2} & \dots & C_{kk} \end{pmatrix}. \tag{4}$$

We may therefore extend CCA to multiple modalities in a very natural way, resulting in the generalized eigenvalue problem

$$\begin{pmatrix} C_{11} & \dots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \dots & C_{kk} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix} = \lambda \begin{pmatrix} C_{11} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & C_{kk} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}. \tag{5}$$

This subsumes two-way CCA as a special case.

2.2. *Kernel Canonical Correlation Analysis*

Kernelization is a principled framework for introducing non-linearity into linear methods. It additionally allows the extension of linear algorithms to non-vectorial domains. For this work, we use the convention that a kernel function is a positive definite, symmetric function that maps two elements of an input space to the real numbers, $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [28]. We denote $\mathcal{H}_x$ the reproducing kernel Hilbert space (RKHS) associated with k_x, and denote the associated feature map $\phi_x : \mathcal{X} \to \mathcal{H}$, i.e. $k_x(x_i, x_j) = \langle \phi_x(x_i), \phi_x(x_j) \rangle$. We note that in general $\phi_x(x_i)$ may no longer have an interpretation in a finite dimensional vector space, but can be viewed as an element in a function space. We analogously define k_y, $\mathcal{H}_y$, and ϕ_y.

We may adapt the representer theorem [29, 28] to the case of multimodal data to state that minimizers of the risk functional

$$\begin{aligned} \min_{f_1,\dots,f_k} \; & c((x_1^1, \dots, x_1^k, f_1(x_1^1), \dots, f_k(x_1^k)), \dots, \\ & (x_n^1, \dots, x_n^k, f_1(x_n^1), \dots, f_k(x_n^k))) + \\ & \sum_{i=1}^k \Omega_k(\|f_i\|^2_{\mathcal{H}_i}), \end{aligned} \tag{6}$$

where c is an arbitrary loss function and Ω a strictly monotonic increasing function, admit representations of the form

$$f_i(x) = \sum_{j=1}^n \alpha_j^i k_i(x_j^i, x), \tag{7}$$

where x_j^i represents the jth sample in the ith modality and $f_i \in \mathcal{H}_i$ a function that maps a sample in the ith modality to the reals. This follows from the representer theorem by considering each modality individually (f_i) while holding all other parameters fixed (f_l where $l \neq i$).

As a result, we may consider a kernelized version of CCA (KCCA). We replace vectors w_i in our previous linear formulation with functions f_i, and replace covariance matrices with the covariance operator

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n}(\phi(x_i) - \mu_\phi)(\phi(x_i) - \mu_\phi)^T, \qquad (8)$$

a linear operator that maps $f \in \mathcal{H}$ to $\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\langle\phi(x_i), f\rangle$ [30]. As we are working with multimodal data, we may consider $\mathcal{H} = \bigoplus_{i=1}^{k}\mathcal{H}_i$ and f to be the concatenation of each f_i. We have used the notation μ_ϕ here to denote the empirical mean of our data sample in the Hilbert space. Analogously to Section 2.1, we may also define cross-covariance and auto-covariance operators $\hat{C}_{xy}$ and $\hat{C}_{xx}$.

Restricting ourselves for the present to the two modality case, we may write the KCCA objective as

$$\max_{f_x, f_y} \frac{f_x^T \hat{C}_{xy} f_y}{\sqrt{f_x^T \hat{C}_{xx} f_x \; f_y^T \hat{C}_{yy} f_y}} = \max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \; \beta^T K_y^2 \beta}}, \qquad (9)$$

where $f_x = \sum_i \alpha_i \phi_x(x_i)$, $f_y = \sum_i \beta_i \phi_y(y_i)$, K_x is the kernel matrix such that $[\tilde{K}_x]_{ij} = k_x(x_i, x_j)$ and $K_x = H\tilde{K}_x H$ where H is a centering matrix

$$H = I - \frac{1}{n}ee^T \qquad (10)$$

$e \in \mathbb{R}^n$ being a vector of all ones. As discussed in Leurgans et al. [2], Bach and Jordan [4], Hardoon et al. [5] this optimization leads to degenerate solutions in the case that either K_x or K_y is invertible so we maximize the following regularized expression

$$\max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T \left(K_x^2 + \varepsilon_x K_x\right)\alpha \beta^T \left(K_y^2 + \varepsilon_y K_y\right)\beta}}, \qquad (11)$$

which is equivalent to Tikhonov regularization of the norms of w_x and w_y in the denominator of Equation (3). In the limit case that $\varepsilon_x \to \infty$ and $\varepsilon_y \to \infty$, the algorithm maximizes covariance instead of correlation.

3. Semi-supervised Kernel Canonical Correlation Analysis

Semi-supervised learning is usually presented in the setting of regression or binary classification [23]. In this setting, the task is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$, where training data are of the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, with additional unlabeled training data available in the $\mathcal{X}$ domain, $\{x_{n+1}, \ldots, x_{n+p_x}\}$.[1] We will use the variable $m_x = n + p_x$ for notational convenience.

3.1. Semi-supervised Laplacian Regularization

Laplacian regularization introduces an additional term into a regularized risk function. One may still regularize using a standard function norm on f, as in Tikhonov regularization, but an additional term penalizes deviations from the data manifold [7]. The representation of the data manifold is estimated empirically from training data, and the additional samples $\{x_{n+1}, \ldots, x_{m_x}\}$ allow us to obtain a much more robust estimate (Figure 2).

In the classic setting, we wish to solve

$$\min_{f_x \in \mathcal{H}_x} c((x_1, y_1, f_x(x_1)), \ldots, (x_n, y_n, f_x(x_n))) \qquad (12)$$
$$+ \varepsilon_x \|f_x\|^2_{\mathcal{H}_x} + \gamma_x \int_{x \in \mathcal{M}} \|\nabla_{\mathcal{M}} f_x\|^2 d\mathcal{P}_x(x)$$

where γ_x is the regularization parameter controlling the degree of Laplacian regularization, $\mathcal{P}_x$ is the marginal distribution of x, and $\nabla_{\mathcal{M}}$ is the gradient of f_x along the manifold $\mathcal{M}$. We do not directly observe $\mathcal{M}$ or $\mathcal{P}_x$ so we must estimate these from the data. As the graph Laplacian converges to the Laplace-Beltrami operator under appropriate conditions [31], we can approximate the integral using the graph Laplacian [7]

$$\min_{f_x \in \mathcal{H}_x} c((x_1, y_1, f_x(x_1)), \ldots, (x_n, y_n, f_x(x_n))) \qquad (13)$$
$$+ \varepsilon_x \|f_x\|^2_{\mathcal{H}_x} + \frac{\gamma_x}{m_x^2} f_x^T \mathcal{L}_{\hat{x}} f_x,$$

where $\hat{x}$ denotes that the empirical graph Laplacian $\mathcal{L}$ was estimated from both labeled and unlabeled data. One may also prove a representer theorem for this form of optimization, in which the minimizer lies in the span of the combined labeled and unlabeled training data [7, Theorem 2]. We extend this here to the case of multimodal data.

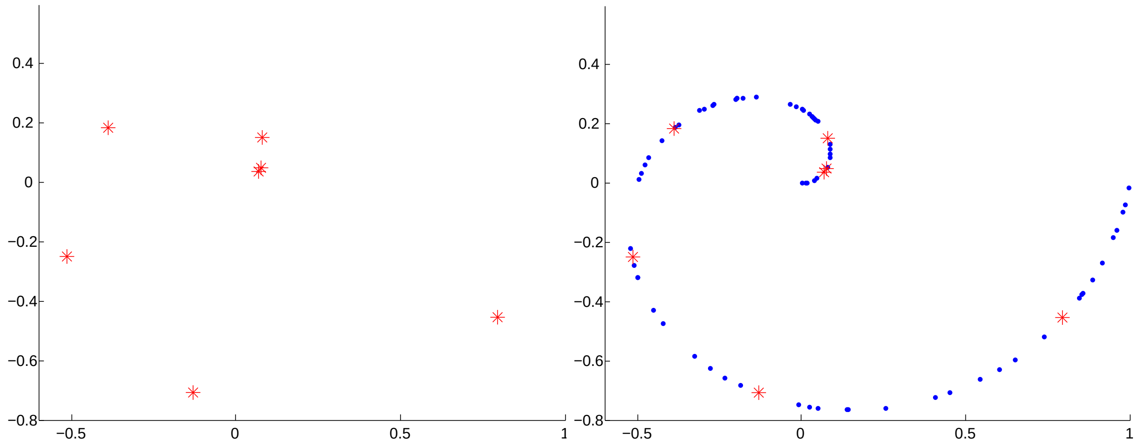
Theorem 1. *Minimizers of the risk functional*

$$\min_{f_1,\ldots,f_k} c((x_1^1, \ldots, x_1^k, f_1(x_1^1), \ldots, f_k(x_1^k)), \ldots,$$
$$(x_n^1, \ldots, x_n^k, f_1(x_n^1), \ldots, f_k(x_n^k))) +$$
$$\sum_{i=1}^{k} \Omega_k(\|f_i\|^2_{\mathcal{H}_i}) + \sum_{i=1}^{k} \frac{\gamma_i}{m_i^2} f_i^T \mathcal{L}_{\hat{x}^i} f_i \qquad (14)$$

admit representations of the form

$$f_i(x) = \sum_{j=1}^{m_i} \alpha_j^i k_i(x_j^i, x) \qquad \forall i. \qquad (15)$$

[1] We return to the setting of data in multiple modalities in Section 3.2.



(a) Labeled data.

(b) Labeled and unlabeled data.

Figure 2: Semi-supervised Laplacian regularization works by employing an additional sample of unlabeled data to improve the estimate of a manifold structure. In (a) only labeled data are shown and the manifold structure is not apparent. In (b) both labeled data (red asterisks) and unlabeled data (blue dots) are shown and the manifold structure is clear.

Proof Fix all modalities arbitrarily except one. Belkin et al. [7, Theorem 2] states that the optimum of this modified problem admits a representation as in Equation (15). As each f_i admits a representation of the form in Equation (15) for arbitrary settings of the other modalities, it also admits such a representation for their optima.

Note that this additionally implies that kernels for each of the modalities can be chosen *independently*.

3.2. The Two-modality Case

We now have the necessary ingredients to apply semi-supervised Laplacian regularization to kernel canonical correlation analysis. KCCA deviates from the classic setting in that modalities $\mathcal{X}$ and $\mathcal{Y}$ are symmetric, and we wish to simultaneously optimize functions that act on each of them. Consequently, we develop notation for kernel matrices with and without semi-supervised data over both the $\mathcal{X}$ and $\mathcal{Y}$ domains. We denote the design matrix $X = (x_1, \dots, x_n)$ where each column represents a data sample that has a correspondence to an observation in $\mathcal{Y}$. We denote the extended design matrix $\hat{X} = (x_1, \dots, x_{m_x})$, in which all data with and without correspondences are stored. We similarly define matrices Y and $\hat{Y}$. We now denote the kernel matrix computed only using the data in X as $K_{xx} \in \mathbb{R}^{n \times n}$, the matrix computed using $\hat{X}$ and X as $K_{\hat{x}x} \in \mathbb{R}^{m_x \times n}$, the matrix computed using $\hat{X}$ with itself as $K_{\hat{x}\hat{x}} \in \mathbb{R}^{m_x \times m_x}$, etc. Kernel matrices for $\mathcal{Y}$ are defined analogously. The following is a semi-supervised Laplacian regularized generalization of Equation (11)

$$\max_{\alpha,\beta} \frac{\alpha^T K_{\hat{x}x} K_{y\hat{y}} \beta}{\sqrt{\alpha^T \left(K_{\hat{x}x} K_{x\hat{x}} + R_{\hat{x}}\right) \alpha \beta^T \left(K_{\hat{y}y} K_{y\hat{y}} + R_{\hat{y}}\right) \beta}}, \tag{16}$$

where $R_{\hat{x}} = \varepsilon_x K_{\hat{x}\hat{x}} + \frac{\gamma_x}{m_x^2} K_{\hat{x}\hat{x}} \mathcal{L}_{\hat{x}} K_{\hat{x}\hat{x}}$ and $R_{\hat{y}} = \varepsilon_y K_{\hat{y}\hat{y}} + \frac{\gamma_y}{m_y^2} K_{\hat{y}\hat{y}} \mathcal{L}_{\hat{y}} K_{\hat{y}\hat{y}}$.

3.3. The General Case

Moving beyond two modalities, we note that the data for which correspondences are known between modalities $\mathcal{X}_i$ and $\mathcal{X}_j$ may be different from the data for which correspondences are known between modalities $\mathcal{X}_j$ and $\mathcal{X}_k$, etc. We abuse the notation $K_{\hat{i}j}$ to denote the kernel matrix computed between all the data for modality i and the data for modality i that also has correspondences to the data in modality j. This matrix has dimensionality $m_i \times n_{ij}$, where m_i is the total number of training examples (with or without correspondences) for modality i, and n_{ij} is the number of correspondences between modalities i and j. The following eigenproblem generalizes Equations (5) and (16)

$$\begin{pmatrix} \mathbf{0} & \dots & \frac{1}{n_{1k}} K_{\hat{1}k} K_{1\hat{k}} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_{1k}} K_{\hat{k}1} K_{k\hat{1}} & \dots & \mathbf{0} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \tag{17}$$

$$\lambda \begin{pmatrix} \frac{1}{m_1} K_{\hat{1}1} K_{1\hat{1}} + R_{\hat{1}} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \frac{1}{m_k} K_{\hat{k}k} K_{k\hat{k}} + R_{\hat{k}} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

3.4. Relationship to Other Subspace Analyses

Principal components analysis (PCA) is a special case of Tikhonov regularized CCA. CCA maximizes a Rayleigh quotient with cross-covariance (C_{xy}) in the numerator, while PCA maximizes a Rayleigh quotient with auto-covariance (C_{xx}) in the numerator. In CCA, when the second modality is simply a copy of the first, the cross-covariance will be equal to the auto-covariance matrix,

$$x_i = y_i \quad \forall i \implies C_{xx} = C_{xy} = C_{yy}. \tag{18}$$

We further note that when $\varepsilon_x \to \infty$ and $\varepsilon_y \to \infty$, the algorithm maximizes cross-covariance instead of cross-correlation. In this case, the maximum of the CCA objective will be achieved when $w_x = w_y$, yielding the direction of greatest auto-covariance, i.e. the solution to PCA.

Additionally, there is an intimate relationship between CCA and Fisher linear discriminant analysis (LDA) [32]. LDA is a special case of CCA where the second modality is the labels [33, 34], consequently, any semi-supervised algorithm for CCA implies a semi-supervised LDA algorithm as well. Recently Cai et al. [24] have proposed a semi-supervised LDA approach. If we use the identity kernel on the labels, set the label regularization parameters to 0, and set $\varepsilon_x = 0$, the directions learned from Equation (16) are the same as those found using the method of Cai et al. [24].

4. Functional Magnetic Resonance Imaging

In neuroscience there has been a recent surge of interest in analyzing brain activity in more natural, complex settings, e.g. with volunteers viewing movies, in order to gain insight in brain processes and connectivity underlying more natural processing. The problem has been approached from different routes: linear regression was used to identify brain areas correlating with particular labels in the movie [35], the perceived content was inferred based on brain activity [36], data-driven methods were used to subdivide the brain into units with distinct response profiles [37], and correlation across subjects was used to infer stimulus-driven brain processes at different timescales [38]. The current approach would allow one to broaden this repertoire to achieve exciting applications. Applied across different brains, the technique may be used to extract commonalities across brains. Finally, weight vectors could e.g. be trained on particular artefacts common to fMRI, such as those induced by heart-rate, breathing or eye-movements, and then could be used to detect related artefacts in novel datasets. With this in mind, we tested the various KCCA variants on data and labels for which the associated and expected brain maps were well known from prior regression analyses [35, 39].

We may formalize this setting as follows. The brain volumes at each time slice may be viewed as training data in the vector space $\mathcal{X}$, where each voxel corresponds to a dimension of the vector space. Associated with the training data are variables measuring the visual content of the stimulus, $\mathcal{Y}$. These may be single or multi-variate. As some of these variables require an expensive manual labeling step, we may optionally include additional training data for which labels are not known.

5. Experimental Results

5.1. Data

fMRI data of one human volunteer was acquired using a Siemens 3T TIM scanner, and consisted of 350 time slices of 3-dimensional fMRI brain volumes. Time-slices were separated by 3.2 seconds (TR), each with a spatial resolution of 46 slices (2.6 mm width, 0.4 mm gap) with 64x64 pixels of 3x3 mm, resulting in a spatial resolution of 3x3x3 mm. The subject watched 2 movies of 18.5 min length, one of which had labels indicating the continuous content of the movie (i.e. degree of visual contrast, or the degree to which a face was present, etc.). The imaging data were pre-processed using standard procedures using the Statistical Parametric Mapping (SPM5) toolbox before analysis [11]. This included a slice-time correction to compensate for acquisition delays between slices, a spatial realignment to correct for small head-movements, a spatial normalization to the SPM standard brain space (near MNI), and spatial smoothing using a Gaussian filter of 6 mm full width at half maximum (FWHM). Subsequently, images were skull-and-eye stripped and the mean of each time-slice was set to the same value (global scaling). A temporal high-pass filter with a cut-off of 512 seconds was applied, as well as a low-pass filter with the temporal properties of the hemodynamic response function (hrf), in order to reduce temporal acquisition noise.

The label time-series were obtained using two separate methods, using computer frame-by-frame analysis of the movie [39], and using subjective ratings averaged across an independent set of five human observers [37]. The computer-derived labels indicated luminance change over time (temporal contrast), visual motion energy (i.e. the fraction of temporal contrast that can be explained by motion in the movie). The

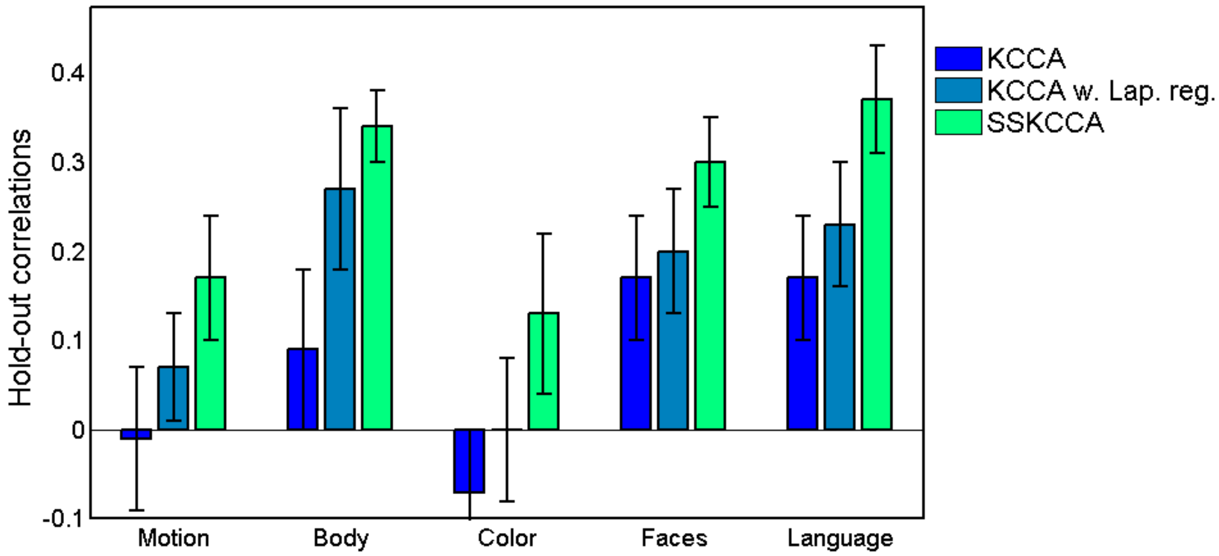


Figure 3: Mean hold-out correlations. From left to right: (i) KCCA makes no use of additional unlabeled data, nor does it model the data manifold; (ii) KCCA with Laplacian regularization does not use unlabeled data, but does use the labeled data to estimate the manifold structure; (iii) Semi-supervised KCCA (SSKCCA) makes use of additional unlabeled samples to better model the data manifold, resulting in increased prediction accuracy.

human-derived labels indicated the intensity of subjectively experienced color, and the degree to which faces and human bodies were present in the movie. In prior studies, each of these labels had been shown to correlate with brain activity in particular and distinct sets of areas specialized to process the particular label in question [37, 39].

5.2. Evaluation Methodology

In order to evaluate the effect of semi-supervised Laplacian regularization on the performance of KCCA, we have evaluated three variants of the algorithm. In the first variant, we have run KCCA without any Laplacian regularization. This is achieved by setting $\gamma_x = \gamma_y = 0$. The second variant consists of Laplacian regularization where the empirical Laplacian matrix was computed using only data for which correspondences between $\mathcal{X}$ and $\mathcal{Y}$ were known. In the final variant, we used full semi-supervised Laplacian regularization, where the manifold was estimated using all available training data. We have not applied Laplacian regularization on the $\mathcal{Y}$ modality in any of the variants, though this may improve performance in that the statistical properties and dependencies of the different image variables may be better modeled. As we are primarily interested in the neuroscientific interpretation of f_x, we have chosen not to exploit these dependencies in this way.

We also evaluate the performance of the algorithms quantitatively. We have run five fold cross validation in which we hold out a portion of the data with correspondences at each fold. As KCCA attempts to maximize Pearson correlation, we first project the held out data using the learned regressors, and then measure their empirical correlation.

In all cases, we have used linear kernels on both the input and output spaces. This is so we may interpret the regressor, f_x, as a learned map of the brain regions implicated in various visual processing. The Laplacian matrix was computed using a Gaussian kernel with the bandwidth parameter set to the median distance between all pairs of training data (with and without correspondences). We have used the symmetric normalized Laplacian $\mathcal{L} = D^{-\frac{1}{2}}(D-W)D^{\frac{1}{2}}$, where D is the diagonal matrix whose entries are the row sums of the similarity matrix, W.

5.3. Model Selection

We have used two model selection criteria to optimize over the variables ε and γ. Both criteria are used as the inner loop of a grid search. In the first variant, we select the model parameters that maximize a five fold cross validation estimate of the empirical correlation (using only the training data). As this is both computationally and statistically inefficient, we have also evaluated a model selection criterion proposed in Hardoon et al.

[5]. This consists of creating a random permutation of the correspondences and running the eigenproblem with the unpermuted data and with the permuted data. The parameter setting with the maximum norm of the difference of the spectra of the two eigenproblems is taken to be the optimum.

5.4. Results

The visual content of the stimulus is quantified in six variables: Motion, Temporal Contrast, Human Body, Color, Faces, and Language. We have repeatedly run all three variants of the experimental setup (Section 5.2) setting our output space to each individual variable. The results for the spectral model selection are shown in Table 1 and Figure 3. We have observed no statistically significant difference when using cross validation, indicating that the spectral model selection criterion can be used in its place. We have additionally run experiments with multi-variate output by grouping several of the variables into three groups: {Visual motion energy, Body, Color}; {Motion, Faces}; and {Motion, Visual motion energy, Color, Faces}. The results of these experiments using the spectral model selection are shown in Table 2.

As we have used linear kernels in all cases, we can interpret the output of the model by analyzing the weights assigned to different spatially localized brain regions. We show results for visual stimulus consisting of *Faces* in Figure 4, *Human body* in Figure 5, *Color* in Figure 6, and *Motion* in Figure 7. In Figure 8 we show results from multivariate output consisting of *Motion* and *Faces*. We further evaluated this case quantitatively, computing the norm of the difference in weight vectors between the multi-variate and univariate cases, which are shown in Table 3. We provide a neuroscientific evaluation in the next section.

6. Discussion

We observe several trends in Tables 1 and 2. First, our major hypotheses were confirmed: for every variate label, the performance improved with the Laplacian regularization on the labeled data, and performance was best in the semi-supervised condition. In the semi-supervised conditions (Experiment 3 as shown in Tables 1 and 2) the additional data without correspondences is sufficiently close to the marginal distribution over $\mathcal{X}$ to improve results significantly, thus the additional data improves the results without any information about the correspondences of the data. Additionally, some variables can be better predicted than others, namely the presence of faces or human bodies in the viewing content, while some elicited relatively poorer performance in all experiments.

Figures 4 through 8 show slices taken through the anatomical image of one subject, with weight maps obtained from the different analyses of its functional data superimposed in red, wherein the maps were thresholded at 2 standard deviations in most cases, but had to be lowered in some cases to reveal any localized activity. Better performance is indicated by more localized activity at a higher threshold, and in all cases there is a clear ranking of the different methods employed. Subfigures 4-8.(a) indicate the complete semi-supervised learning framework and show the most localized regions, few red areas fall outside of the ovals indicating the regions of expected activity. Subfigures 4-8.(b) indicate the use of Laplacian regularization, but without the inclusion of unlabeled data and have slightly less pronounced weights where expected (white ovals) and somewhat more spurious areas of high weights in other regions. Finally, Subfigures 4-8.(c) show the results of KCCA without Laplacian regularization. This learning framework was considered state-of-the-art prior to the present work [17], but shows severely degraded performance: few regions of high weight fall within the expected regions (white ovals) while large amounts of spurious activity is present.

We show examples of four of the single-variate labels for each of the three experiments, as well as one of the sets of multi-variate experiments. In the multi-variate label example, we show the same weight map but at different brain volume coordinates in order to visualize the expected brain activations for each of the lables involved. Summary statistics of the relationship between the multi-variate label example and the univariate weights are given in Table 3, indicating that the solution was overall much closer to that of Faces, which can be predicted much more reliably (Table 1). The maps correspond well to the known functional anatomy, and to activations obtained in the previous regression studies of free-movie-viewing data [37]. Faces obtained high weights in the fusiform cortex (fusiform face area, FFA) (Figure 4); Human Bodies dorso-lateral and ventral parts within the lateral occipital cortex (extrastriate body area (EBA) and fusiform body area (FBA)) (Figure 5); Color obtained high weights in the medial fusiform cortex where human V4 is located (Figure 6). The spatial layout of the weights thus corresponds well to the previous literature, and indicates that some of the analyses applied here yield results that are neuroscientifically meaningful and that can identify distinct cortical regions involved in the distinct tasks. Semi-

Table 1: Mean holdout correlations across the six variables in all experiments with the spectral model selection criterion of Hardoon et al. [5]. Experiment 1 consists of KCCA using only data for which correspondences are known. Experiment 2 employs Laplacian regularization where the Laplacian matrix is estimated using only data for which correspondences are known. Finally, experiment 3 employs full semi-supervised Laplacian regularization. Semi-supervised Laplacian regluarization gives the best performance in all cases.

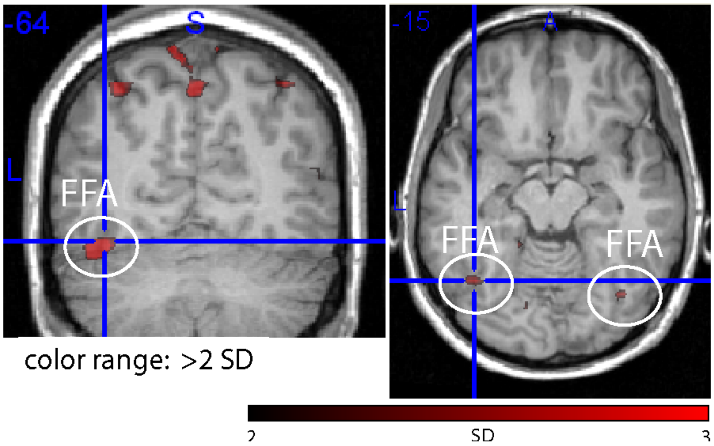
	Exp 1	Exp 2	Exp 3
Motion	-0.012 ± 0.081	0.065 ± 0.066	**0.170** ± **0.074**
Temporal Contrast	0.042 ± 0.065	0.088 ± 0.084	**0.116** ± **0.101**
Human Body	0.095 ± 0.086	0.274 ± 0.093	**0.340** ± **0.043**
Color	-0.075 ± 0.069	-0.002 ± 0.079	**0.128** ± **0.089**
Faces	0.173 ± 0.073	0.203 ± 0.075	**0.303** ± **0.054**
Language	0.172 ± 0.070	0.231 ± 0.074	**0.365** ± **0.057**

Table 2: Mean holdout correlations across the 3 multi-variate sets in all experiments with the spectral model selection criterion of Hardoon et al. [5]. Experiment 1 consists of KCCA using only data for which correspondences are known. Experiment 2 employs Laplacian regularization where the Laplacian matrix is estimated using only data for which correspondences are known. Finally, experiment 3 employs full semi-supervised Laplacian regularization. Semi-supervised Laplacian regluarization gives the best performance in all cases.

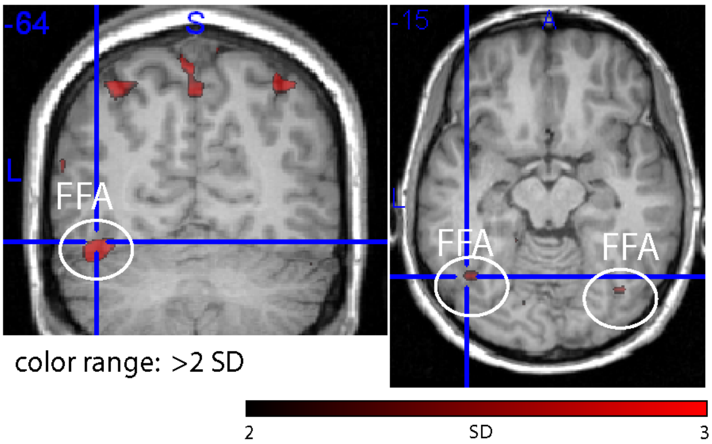
	Exp 1	Exp 2	Exp 3
{Visual motion energy, Body, Color}	0.1596 ± 0.0807	0.1873 ± 0.0879	**0.2844** ± **0.0716**
{Motion, Faces}	-0.0827 ± 0.0460	0.0602 ± 0.0908	**0.1898** ± **0.0636**
{Motion, Visual motion energy, Color, Faces}	0.1167 ± 0.0785	0.1498 ± 0.0827	**0.2528** ± **0.0579**

Table 3: Norms of the difference of weight vectors for Motion (w_a), Faces (w_b), and {Motion, Faces} (w_c) experiments. Weight vectors are computed using semi-supervised Laplacian regularization in all cases. Portions of the weight maps are visualized in Figures 4, 7, and 8.

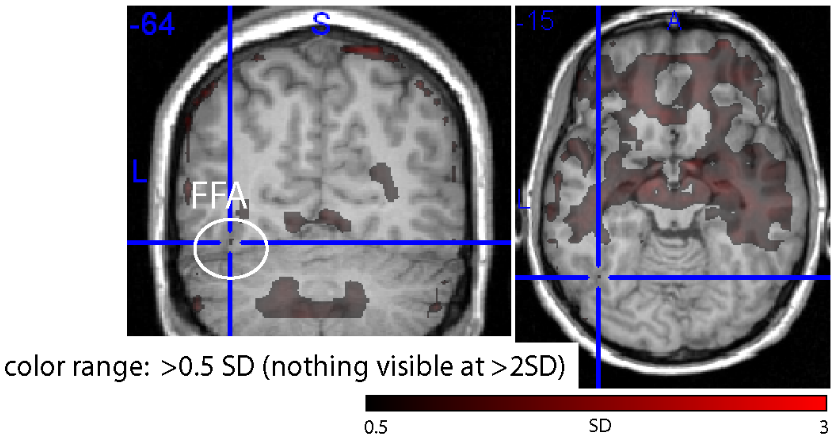
$\|\|w_a - w_b\|\|$	$\|\|w_c - w_a\|\|$	$\|\|w_c - w_b\|\|$
1.37	1.29	0.10



(a) Semi-supervised Laplacian regularized solution.

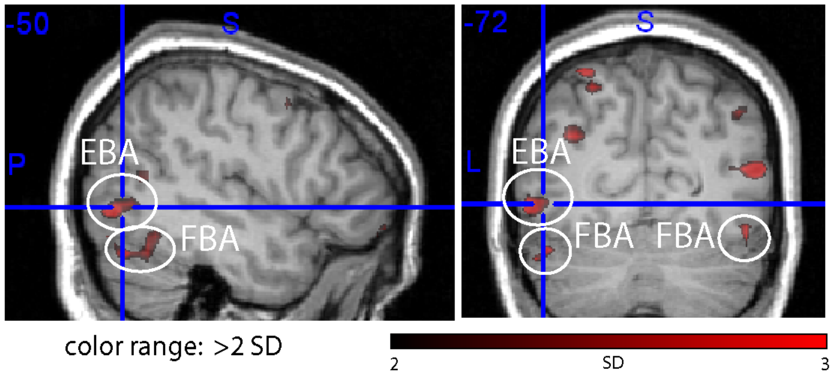


(b) Laplacian regularized solution.

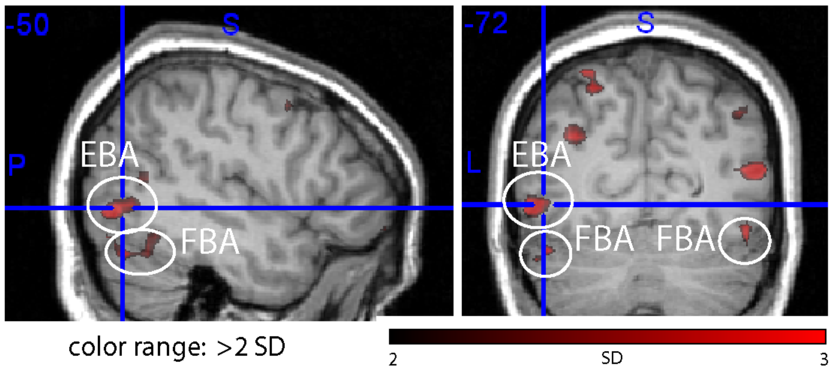


(c) KCCA without Laplacian regularization.

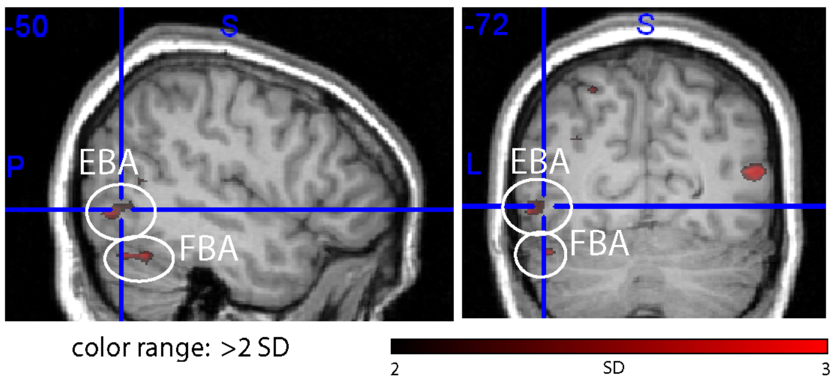
Figure 4: Faces: activation in the cortical region responsive to the visual perception of faces, the fusiform face area (FFA). Weight vectors are plotted over an anatomical image of the volunteers brain. Note that the semi-supervised Laplacian regularization led to the most specific and most significant weights in FFA.



(a) Semi-supervised Laplacian regularized solution.

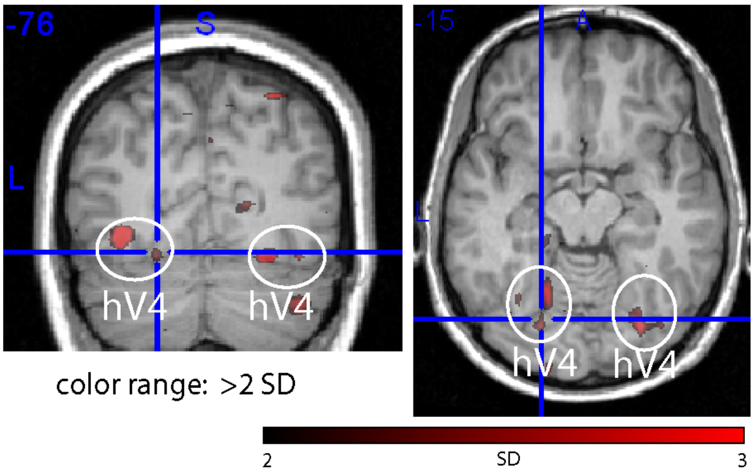


(b) Laplacian regularized solution.

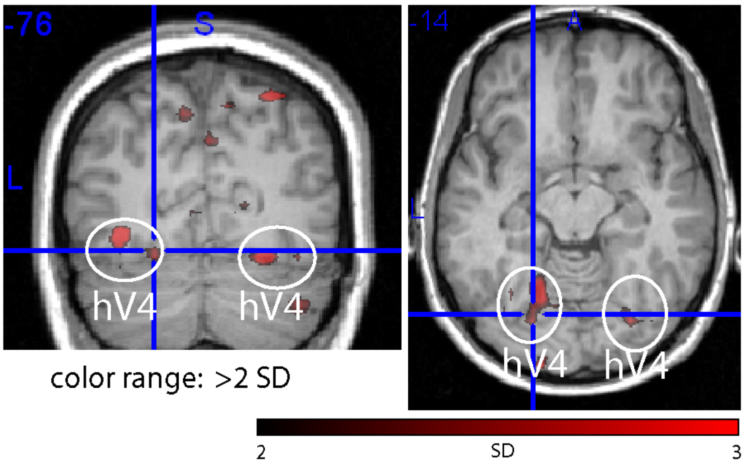


(c) KCCA without Laplacian regularization.

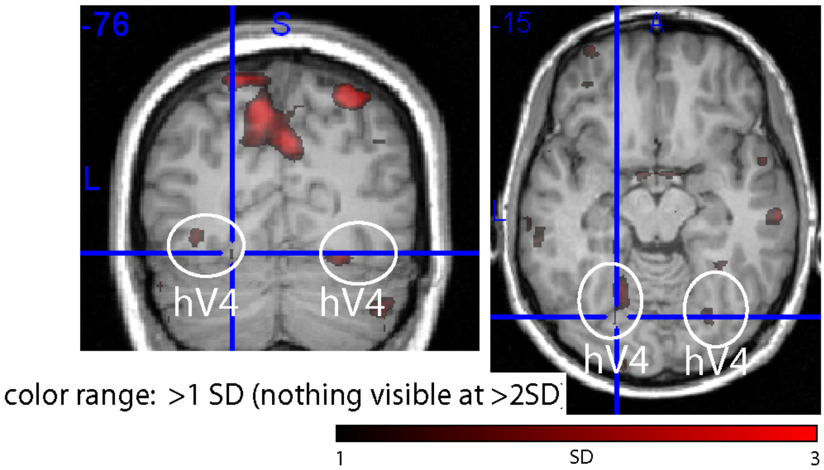
Figure 5: Human Body: activation in the cortical region responsive to the visual perception of human bodies, in the extrastriate body area (EBA) and in the fusiform body area (FBA). Same observation as in Figure 4.



(a) Semi-supervised Laplacian regularized solution.

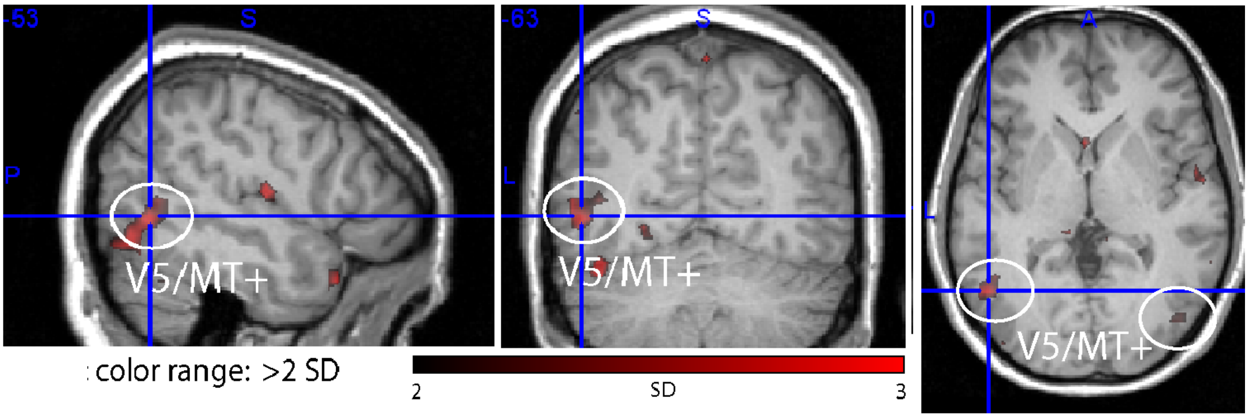


(b) Laplacian regularized solution.

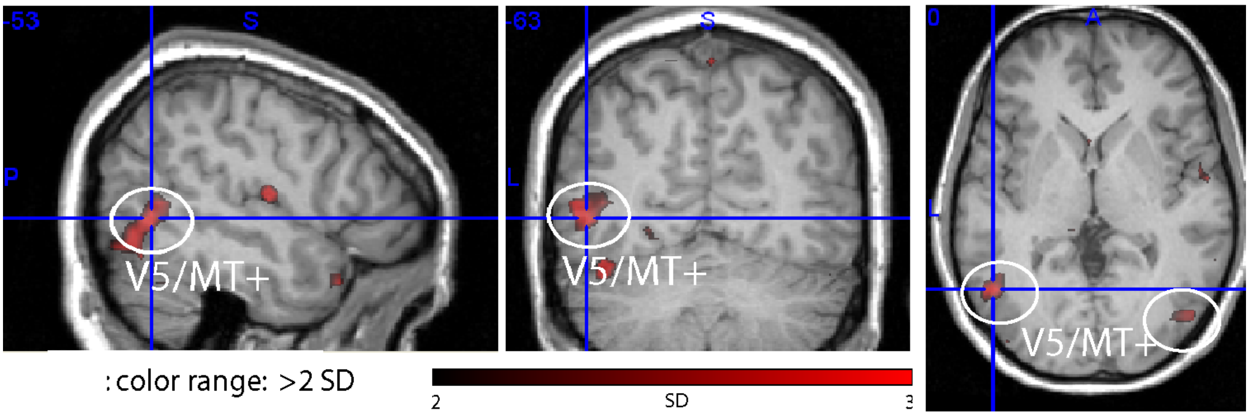


(c) KCCA without Laplacian regularization.

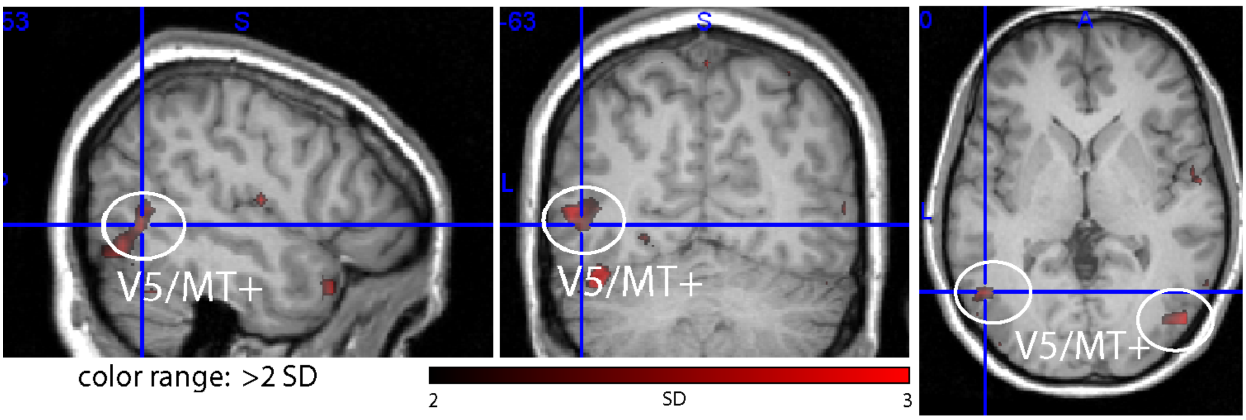
Figure 6: Color: activation in the color responsive cortex (human visual area 4, hV4). Same observation as in Figure 4.



(a) Semi-supervised Laplacian regularized solution.

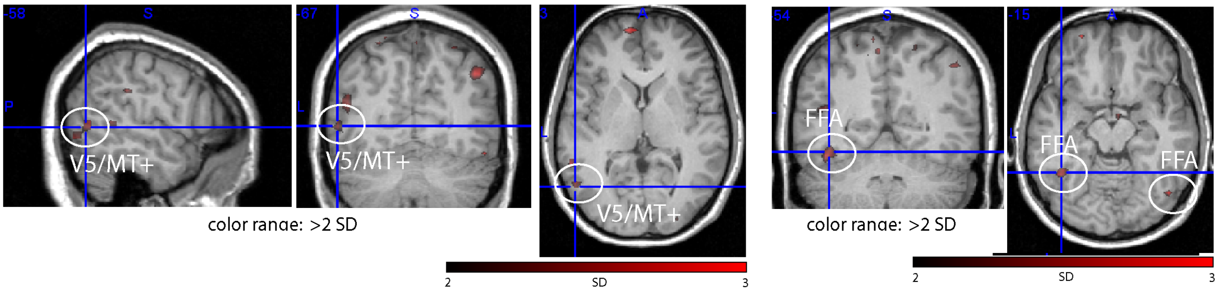


(b) Laplacian regularized solution.

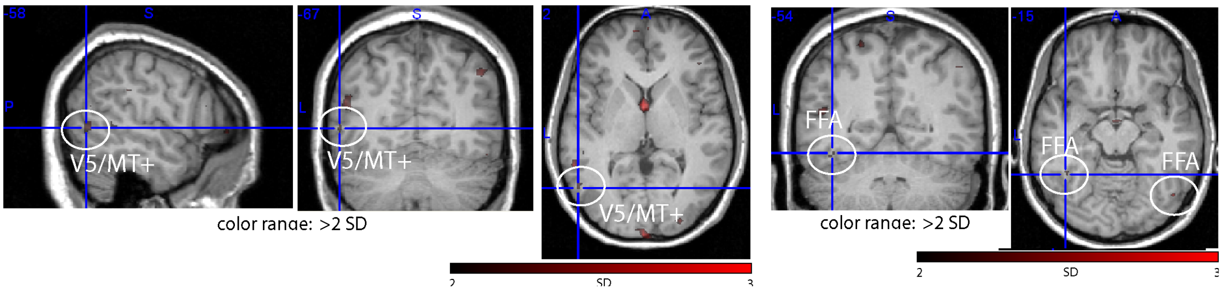


(c) KCCA without Laplacian regularization.

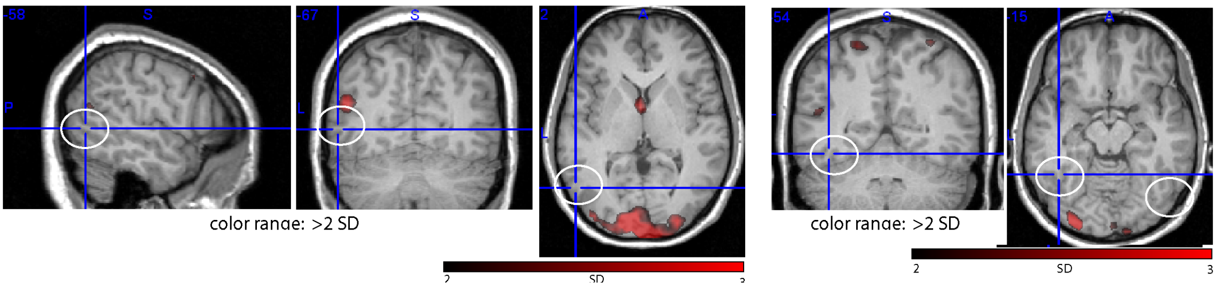
Figure 7: Motion: activation in the visual motion complex, area V5+/MT+. Same observation as in Figure 4.



(a) Semi-supervised Laplacian regularized solution.



(b) Laplacian regularized solution.



(c) KCCA without Laplacian regularization.

Figure 8: Multivariate - *Motion* and *Faces*: activations in the visual motion complex, area V5+/MT+ (left), and activation in the cortical region responsive to the visual perception of faces, the fusiform face area (FFA) (right). Same observation as in Figure 4.

supervised Laplacian regularization worked well in that weight maps thresholded at >2SD show relatively well defined activity of the regions previously shown to be involved with the features. For other analyses, e.g. KCCA without Laplacian regularization, we had to reduce the threshold to 0.5 or 1 (faces and color in the single-variate cases, respectively) to obtain activity in the areas in question, and the maps show additional, unspecific activity as well.

7. Conclusions and Future Work

In this work, we used a semi-supervised Laplacian regularized generalization of KCCA which has several important regression techniques as special cases. The various experimental designs we tested improved successively (as shown by the correlations of the holdout sets for each variant), both from the supervised variant with no regularization parameter, to the supervised variant with Laplacian regularization, and further improved in the semi-supervised variant by the addition of unlabeled data. Additionally, the analysis of the weights learned by, particularly the semi-supervised experiment and the supervised Laplacian regularized experiment, display the same brain activation patterns shown in previous neuroscientific studies. This provides promise that with semi-supervised methods, one can simply add unlabeled data of a similar variety to a (significantly smaller) labeled data set, and achieve similar results, as the approximation of the labeled data distribution is thereby strengthened. These results additionally lay the groundwork for exciting neuroscience applications, such as removal of learned artefacts from fMRI data, eliminating the need for expensive labeling of natural stimuli shown during fMRI image acquisition, as well as potential for discovering brain activity patterns associated to new or unknown stimuli.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007- 2013) / ERC grant agreement no. 228180. This work was funded in part by the EC project CLASS, IST 027978, and the PASCAL2 network of excellence, IST 2002-506778. MBB is funded by the Royal Academy of Engineering through a Newton International Fellowship. This work was done when Arthur Gretton was resident at the Max Planck Institute for Biological Cybernetics. We acknowledge partial funding by the project BMBF 01GQ0840 (BFNT Frankfurt).

References

[1] H. Hotelling, Relations Between Two Sets of Variates, Biometrika 28 (1936) 321–377.

[2] S. E. Leurgans, R. A. Moyeed, B. W. Silverman, Canonical Correlation Analysis when the Data are Curves, Journal of the Royal Statistical Society, Series B (Methodological) 55 (3) (1993) 725–740.

[3] P. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, Int. J. Neural Syst. 10 (5) (2000) 365–377.

[4] F. R. Bach, M. I. Jordan, Kernel Independent Component Analysis, JMLR 3 (2002) 1–48.

[5] D. R. Hardoon, S. Szedmák, J. R. Shawe-Taylor, Canonical Correlation Analysis: An Overview with Application to Learning Methods, Neural Computation 16 (12) (2004) 2639–2664, ISSN 0899-7667, doi:http://dx.doi.org/10.1162/0899766042321814.

[6] M. B. Blaschko, C. H. Lampert, Correlational Spectral Clustering, in: CVPR, 2008.

[7] M. Belkin, P. Niyogi, V. Sindhwani, Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples, JMLR 7 (2006) 2399–2434.

[8] M. Turk, A. Pentland, Face recognition using eigenfaces, 586–591, doi:10.1109/CVPR.1991.139758, 1991.

[9] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neuroscience 3 (1) (1991) 71–86, ISSN 0898-929X, doi: http://dx.doi.org/10.1162/jocn.1991.3.1.71.

[10] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, Pattern Analysis and Machine Intelligence, IEEE Transactions on 19 (7) (1997) 711–720, ISSN 0162-8828, doi:10.1109/34.598228.

[11] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, W. Penny (Eds.), Statistical Parametric Mapping: The Analysis of Functional Brain Images, Academic Press, 2007.

[12] R. Chaves, J. Ramírez, J. Górriz, M. López, D. Salas-Gonzalez, I. Álvarez, F. Segovia, SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting, Neuroscience Letters 461 (3) (2009) 293–297, ISSN 0304-3940, doi: 10.1016/j.neulet.2009.06.052.

[13] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex, Science 293 (5539) (2001) 2425–2430, doi:10.1126/science.1063736.

[14] J.-D. Haynes, G. Rees, Decoding mental states from brain activity in humans, Nature Reviews Neuroscience 7 (7) (2006) 523–534, ISSN 1471-003X, doi:10.1038/nrn1931, URL `http://dx.doi.org/10.1038/nrn1931`.

[15] S.-P. Ku, A. Gretton, J. Macke, N. K. Logothetis, Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys, Magnetic Resonance Imaging 26 (7) (2008) 1007 – 1014, ISSN 0730-725X, doi:DOI: 10.1016/j.mri.2008.02.016.

[16] J. Ramírez, J. Górriz, R. Chaves, M. López, D. Salas-Gonzalez, I. Álvarez, F. Segovia, SPECT image classification using random forests, Electronics Letters 45 (12) (2009) 604–605, doi: 10.1049/el.2009.1111.

[17] D. R. Hardoon, J. Mourão-Miranda, M. Brammer, J. Shawe-Taylor, Unsupervised Analysis of fMRI Data Using Kernel Canonical Correlation, NeuroImage 37 (4) (2007) 1250–1259.

[18] Y. Li, J. Shawe-Taylor, Using KCCA for Japanese—English cross-language information retrieval and document classification, J. Intell. Inf. Syst. 27 (2) (2006) 117–133, ISSN 0925-9902, doi:http://dx.doi.org/10.1007/s10844-006-1627-y.

[19] Y. Yamanishi, J.-P. Vert, A. Nakaya, M. Kanehisa, Extraction of correlated gene clusters from multiple genomic

data by generalized kernel canonical correlation analysis, Bioinformatics 19 (suppl 1) (2003) i323–330, doi: 10.1093/bioinformatics/btg1045.
[20] J. Dauxois, G. M. Nkiet, Nonlinear Canonical Analysis and Independence Tests, Ann. Statist. 26 (4) (1998) 1254–1278.
[21] K. Fukumizu, A. Gretton, X. Sun, B. Schölkopf, Kernel Measures of Conditional Dependence, in: NIPS, 2007.
[22] M. B. Blaschko, C. H. Lampert, A. Gretton, Semi-supervised Laplacian Regularization of Kernel Canonical Correlation Analysis, in: ECML PKDD '08: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I, Springer-Verlag, Berlin, Heidelberg, ISBN 978-3-540-87478-2, 133–145, 2008.
[23] O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT Press, Cambridge, MA, URL `http://www.kyb.tuebingen.mpg.de/ssl-book`, 2006.
[24] D. Cai, X. He, J. Han, Semi-Supervised Discriminant Analysis, in: ICCV, 2007.
[25] M. B. Blaschko, J. A. Shelton, A. Bartels, Augmenting Feature-driven fMRI Analyses: Semi-supervised Learning and Resting State Activity, in: Advances in Neural Information Processing Systems, 2009.
[26] V. R. de Sa, P. W. Gallagher, L. J. M., V. L. Malave, Multi-view kernel construction, Machine Learning .
[27] F. Bießmann, F. C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. K. Logothetis, K.-R. Müller, Temporal kernel CCA and its application in multimodal neuronal data analysis, Mach. Learn. 79 (1-2) (2010) 5–27.
[28] B. Schölkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, The MIT Press, ISBN 0262194759, 2002.
[29] G. S. Kimeldorf, G. Wahba, Some Results on Tchebycheffian Spline Functions, Journal of Mathematical Analysis and Applications 33 (1) (1971) 82–95.
[30] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Neural Computation 10 (5) (1998) 1299–1319, ISSN 0899-7667, doi: http://dx.doi.org/10.1162/089976698300017467.
[31] M. Hein, J.-Y. Audibert, U. v. Luxburg, Graph Laplacians and their Convergence on Random Neighborhood Graphs, J. Mach. Learn. Res. 8 (2007) 1325–1370, ISSN 1533-7928.
[32] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.
[33] T. De Bie, Semi-supervised learning based on kernel methods and graph cut algorithms, Phd thesis, K.U.Leuven (Leuven, Belgium), Faculty of Engineering, 2005.
[34] F. R. Bach, M. I. Jordan, A Probabilistic Interpretation of Canonical Correlation Analysis, Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.
[35] A. Bartels, S. Zeki, Functional brain mapping during free viewing of natural scenes., Human Brain Mapping 21 (2) (2004) 75–85.
[36] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach, Intersubject Synchronization of Cortical Activity During Natural Vision, Science 303 (5664) (2004) 1634–1640, doi: 10.1126/science.1089506.
[37] A. Bartels, S. Zeki, The chronoarchitecture of the human brain–natural viewing conditions reveal a time-based anatomy of the brain, NeuroImage 22 (1) (2004) 419 – 433, ISSN 1053-8119, doi:DOI: 10.1016/j.neuroimage.2004.01.007.
[38] U. Hasson, E. Yang, I. Vallines, D. J. Heeger, N. Rubin, A Hierarchy of Temporal Receptive Windows in Human Cortex, J. Neurosci. 28 (10) (2008) 2539–2550, doi:10.1523/JNEUROSCI.5487-07.2008, URL `http://dx.doi.org/10.1523/JNEUROSCI.5487-07.2008`.
[39] A. Bartels, S. Zeki, N. K. Logothetis, Natural Vision Reveals Regional Specialization to Local Motion and to Contrast-Invariant, Global Flow in the Human Brain, Cereb. Cortex (2007) bhm107+doi:10.1093/cercor/bhm107, URL `http://dx.doi.org/10.1093/cercor/bhm107`.